Predicting Peptides Binding to MHC Class II Molecules Using Multi-objective Evolutionary Algorithms

BMC Bioinformatics

BioMed Central

Research article

Predicting peptides binding to MHC class II molecules using multi-objective evolutionary algorithms

Menaka Rajapakse*[1,3], Bertil Schmidt[2], Lin Feng[3] and Vladimir Brusic[4]

Address: [1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, 119613 Singapore, [2]NICTA VRL, University of Melbourne, Parkville, 3010 Australia, [3]School of Computer Engineering, Nanyang Technological University, Block N4, Nanyang Avenue, 639798 Singapore and [4]Cancer Vaccine Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115 USA

Email: Menaka Rajapakse* - menaka@i2r.a-star.edu.sg; Bertil Schmidt - bertil.schmidt@computer.org; Lin Feng - asflin@ntu.edu.sg; Vladimir Brusic - vladimir_brusic@dfci.harvard.edu

* Corresponding author

Published: 22 November 2007

BMC Bioinformatics 2007, **8**:459 doi:10.1186/1471-2105-8-459

Received: 7 May 2007
Accepted: 22 November 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/459

Abstract

Background: Peptides binding to Major Histocompatibility Complex (MHC) class II molecules are crucial for initiation and regulation of immune responses. Predicting peptides that bind to a specific MHC molecule plays an important role in determining potential candidates for vaccines. The binding groove in class II MHC is open at both ends, allowing peptides longer than 9-mer to bind. Finding the consensus motif facilitating the binding of peptides to a MHC class II molecule is difficult because of different lengths of binding peptides and varying location of 9-mer binding core. The level of difficulty increases when the molecule is promiscuous and binds to a large number of low affinity peptides.

In this paper, we propose two approaches using multi-objective evolutionary algorithms (MOEA) for predicting peptides binding to MHC class II molecules. One uses the information from both binders and non-binders for self-discovery of motifs. The other, in addition, uses information from experimentally determined motifs for guided-discovery of motifs.

Results: The proposed methods are intended for finding peptides binding to MHC class II I-A^{g7} molecule – a promiscuous binder to a large number of low affinity peptides. Cross-validation results across experiments on two motifs derived for I-A^{g7} datasets demonstrate better generalization abilities and accuracies of the present method over earlier approaches. Further, the proposed method was validated and compared on two publicly available benchmark datasets: (1) an ensemble of qualitative HLA-DRB1*0401 peptide data obtained from five different sources, and (2) quantitative peptide data obtained for sixteen different alleles comprising of three mouse alleles and thirteen HLA alleles. The proposed method outperformed earlier methods on most datasets, indicating that it is well suited for finding peptides binding to MHC class II molecules.

Conclusion: We present two MOEA-based algorithms for finding motifs, one for self-discovery and the other for guided-discovery by experimentally determined motifs, and thereby predicting binding peptides to I-A^{g7} molecule. Our experiments show that the proposed MOEA-based algorithms are better than earlier methods in predicting binding sites not only on I-A^{g7} but also on most alleles of class II MHC benchmark datasets. This shows that our methods could be applicable to find binding motifs in a wide range of alleles.

Background

Major histocompatibility complex (MHC) molecules play a key role in initiating immune responses. They bind to and expose an antigen (or short peptides) to T cell receptors (TCR) triggering an immune response against the infected cell or foreign agent. MHC molecules make multiple contacts with the side-chains of binding peptides, which define the binding motif and determine the specificity of binding [1]. Prediction of peptides binding to a MHC class II molecule is difficult due to different types of side chains and because the length of the binding peptides is longer than 9aa (approximately 11 to 22aa) [1,2]. It has been previously observed that a core of 9aa is sufficient for binding peptides to a MHC class II molecules [3], however, the exact location of the binding core (or motif) within the peptide is usually unknown and vary.

A binding motif is usually represented either by a consensus sequence or as a weight matrix [4]. The presence or composition of a motif can be experimentally determined from a large pool of putative binding peptides [3,5]. However, such wet-lab experiments are costly, time consuming, and cumbersome. Amino acids at specific sites of a motif, contributing significantly to the binding are referred to as *primary anchor residues* and the corresponding sites as *anchor positions*. By using such position-specific information, earlier studies have found weight matrix models elaborating the nature and strength of binding motifs [6,7]. These models offer binding strengths of every residue at specific sites in the form of a position specific scoring matrix (PSSM).[7]

In general, MHC class-II prediction methods are categorized into two main classes [8]: (1) quantitative prediction methods that predict inhibitory concentration (IC_{50}) values and (2) qualitative prediction methods that determine the binding status (binder or non-binder) based on the predictive score. Recent quantitative prediction approaches include SVRMHC [8], PLS-ISC [9], ARB [10], and SMM-align [11]. The ARB approach uses full length of the peptide whereas both SVRMHC and PLS-ISC approaches use a preprocessing step involving alignment of sequences, based on anchor position-specific residues. The underlying assumption of SMM-align is that amino acids occupying the 9-mer binding core motif are sufficient to determine the affinity of peptide-MHC binding. However, in some cases, the predictive performance could be improved by incorporating terminal residues known as peptide flanking residues (PFR) [11].

Qualitative prediction approaches use classifiers such as artificial neural networks [12-16], hidden Markov models [4,17], support vector machines [18-21], and their hybrids [22], or profile analysis such as those using iterative learning [23-26], stochastic approaches (MEME) [27,28], Gibbs motif sampler [29-32], profile motifs (RANKPEP) [33,34], DNA microarrays and virtual matrices (TEPITOPE) [35], and evolutionary algorithms (EA) [36]. However, given a set of sequences of differing lengths with known binding affinities, the location of the binding core within each sequence must be first identified before classification of sequences. Classical multiple sequence alignment techniques often fail to detect binding cores in MHC class II binding peptides because of weak instances of binding motifs.

All methods predicting peptides binding to MHC molecules have their pros and cons; most show good performance only for datasets upon which they were developed. Therefore, there is a need for new algorithms that perform well on previously unseen data. We propose to use MOEA to align a set of experimentally determined binding peptides at their binding cores and subsequently derive the consensus motif. The methods are especially useful when molecules are promiscuous and bind to a large number of low affinity peptides. The preliminary results of our work have been presented in [37].

I-A^{g7} is the MHC class II molecule of the NOD mouse, critical for the development of insulin-dependent diabetes mellitus (IDDM) and other autoimmune disorders [38-43]. Knowledge of peptides binding to I-A^{g7} is important in understanding the molecular basis of development of IDDM in NOD mice. Experiments have demonstrated that I-A^{g7} binding peptides are 9–30aa long [44]. Finding motifs in peptide binding to I-A^{g7} is a non-trivial problem [45,46]. Despite numerous attempts, no consensus has been reached on the rules of peptide binding to I-A^{g7} molecule [38-48]. However, computational analyses on multiple datasets indicate that experimental motifs satisfy only a subset of rules describing the optimal motif.

To demonstrate the utility in predicting peptides binding to other MHC molecules, our method is tested on two benchmark datasets comprising of peptides of number of different HLA (human MHC) and mouse alleles. The first dataset, referred to as BM-Set1 here onwards, consists of different combinations of peptides of HLA-DRB1*0401 allele, and the second dataset, BM-Set2, consists of datasets from thirteen different HLA alleles and three mouse alleles.

Multi-Objective Evolutionary Algorithms (MOEA)

Evolutionary algorithms (EA) are based on the principles of biological evolution and have often been successful in solving complex search and optimization problems. Majority of bioinformatics applications of EA have been in the discovery of motifs such as transcription factor binding sites [49-53]. Yet, only a few researchers have

used EA for the prediction of peptides binding to protein sequences [36].

An EA consists of (1) representing input variables as individuals or chromosomes (binary or real valued) in a population, (2) formulating the fitness (objective function) to evaluate individuals, (3) generating a new population by genetic operations (such as reproduction, crossover, and mutation) on the current population, and (4) determining if the population has reached the optimal fitness. The algorithm begins with an initial population and evolves over time. At a particular instance of evolution, every individual is evaluated by its fitness. New populations (offspring) are produced from highly fit individuals (parents) selected, which undergo genetic operations. Each offspring is paired and compared to its parents. Highly fit individuals are retained in the population while less fit individuals are discarded. Search mechanisms such as elitism, constraint-handling, and multi-objective optimization are available for finding a better spread of solutions, depending on the needs of the optimization problem [54-57].

Multi-objective evolutionary algorithms (MOEA) are used to solve problems which require simultaneous optimization of a number of competing objective functions [58-61]. MOEA maintains a set of solutions ranked by their dominance at a given instant of the evolution. A solution is said to dominate another if it is better or equal with respect to all objectives and strictly better in at least one objective [58]. Often, there are more than one non-dominated solutions, representing the best ones, collectively known as the *Pareto* front. MOEA algorithms result in a *Pareto optimal set* of solutions.

Non-dominated Sorting Genetic Algorithm II (NSGA-II) was recently introduced to incorporate several new genetic mechanisms for better convergence, such as non-dominated sorting, elitism, diversity preservation, and constraint handling [58]. In NSGA-II, a population is subjected to several rounds of non-dominated sorting. That is, all the non-dominated individuals are identified and assigned the same fitness value until a new set of non-dominated solutions is found. The solutions found in subsequent rounds are assigned fitness values lower than those in the previous rounds. This process continues until the whole population is partitioned into non-dominated fronts with diverse fitness values. The elitism prevents the loss of fit individuals encountered in earlier generations by allowing earlier solutions to survive in the subsequent generations. The diversity of Pareto-optimal solutions is maintained by imposing a measure referred to as *crowding distance*. A solution that satisfies the constraints defined by the objective functions is called a *feasible solution*.

Peptide Binding to MHC Class II I-A^{g7}

In this paper, we attempt to find an optimal motif describing peptide binding to MHC class II molecules, using experimentally determined binding data. There are several factors that impede the derivation of such a consensus motif. The first is the strong resemblance among the peptides isolated in a single experiment and the second is the diversity among different datasets. A motif derived from a dataset lacking diversity indicates a bias towards the dataset used in deriving the motif. Such motifs are difficult to generalize on other experimental or previously unseen datasets. The MOEA based motif detection algorithm is designed to find a consensus motif on I-A^{g7} datasets, which alleviates the influences arising from biased datasets and thereby predicts binding peptides more accurately in new datasets.

Results

Predicting Peptides Binding to MHC Class II

We use our approach to find a consensus motif on seven experimental datasets of peptides binding to I-A^{g7} molecules, obtained from literature [40-43,62-64]. The motif is validated using an independent testing set generated from the Stratmann dataset [46]. The overall quality of prediction was measured using area under curve (AUC) of the receiver operating characteristics (ROC) curve [65-67]. AUC values of all feasible solutions in the final population of EA were evaluated and the solution with the highest AUC was chosen as the consensus motif (see Additional file 1).

Table 1 shows the information of the datasets extracted from literature, which were used in the training. A blank '-' indicates the unavailability of a particular information. As an example, the details of the experimental motif of Reizis *et al* are given in Table 2. Table 3 shows the performance when an experimental motif is used to predict peptide binders in other datasets. As seen, a motif of a particular experiment does not characterize peptide binding of I-A^{g7} molecules in other datasets. Table 4 shows the cross-validation performance of two motifs (by self-discovery and guided-discovery) derived using MOEA; in a particular cross-validation run, one experimental dataset was excluded and the motif was derived using the information of the remaining datasets. The motif was tested for predicting binders and non-binders of the left-out dataset. The self-discovery approach uses only the binding information whereas the guided-discovery uses both binding information as well as information associated with experimental motifs. As seen in Table 4, by achieving AUC values greater than 0.7 for all cross-validation runs, MOEA derived motifs demonstrate better generalization capabilities compared to experimentally determined motifs. The binding motifs derived from self-discovery and guided-

Table 1: I-A^{g7} datasets and experimental motifs

Dataset	Experimental Motif	Non-binders	Binders	Reference
Reizis	*m*(Reizis)	21	33	[40]
Harrison	*m*(Harrison)	19	157	[41]
Gregori	*m*(Gregori)	31	109	[43]
Latek	*m*(Latek)	8	37	[42]
-	*m*(Rammensee)	-	-	[44]
-	*m*(Reich)	-	-	[38]
-	*m*(Amor)	-	-	[39]
Corper	-	35	13	[62]
MHCPEP	-	-	176	[63]
Yu	-	16	10	[64]
Brusic	-	37	-	[unpublished]

Information on I-A^{g7} related peptide binding datasets and motifs. Unavailable information is indicated by "-".

discovery are illustrated as sequence logo plots [68] in the Additional file 2.

To compare the performance of our method with earlier methods, a training dataset was created by combining all the experimental datasets given in Table 1. Motifs derived on the training dataset were tested on an independent test dataset – a balanced set generated from Stratmann dataset. The Stratmann dataset was balanced by adding randomly generated non-binders. Twenty five such balanced test datasets were assembled by generating random samples starting from different seeds and adding them to the Stratmann dataset. The results reported are based on the average AUC values over all balanced test sets. Figure 1 shows comparison of performances of motifs derived by MOEA and by earlier motif prediction approaches such as MEME and RANKPEP. An increase of 4–10% in predictive performance is observed with MOEA over the other approaches.

Table 2: Representation of an experimentally derived I-A^{g7} motif

Position	Well-Tolerated	Weakly-Tolerated	Non-Tolerated
P1	VEQMHLPD	-	R
P2	-	-	-
P3	-	-	-
P4	**ILPV**	**HY**	**QEK**
P5	-	-	-
P6	**ATSNV**	**-**	**LYQK**
P7	QVYLHINRF	-	-
P8	-	-	-
P9	**ED**	**SM**	**LYTQK**

The description of experimentally determined I-A^{g7} 9-mer peptide binding motif by Reizis: each position accommodates a well-tolerated, weakly-tolerated, or non-tolerated amino acid. The positions P4, P6 and P9 are the primary anchor positions where binding is highly likely to occur.

Comparison of performances of MOEA derived motifs for BM-Set1 (see Table 5) with enhanced Gibbs sampler [32], TEPITOPE [35], SVRMHC [8] and ARB [10], is given in Table 6. As seen, MOEA shows comparable or superior performance with Gibbs sampler on all datasets except for the Southwood dataset. Out of the ten non-redundant (NR) datasets, the MOEA outperformed Gibbs sampler, TEPITOPE, SVRMHC and ARB by seven, nine, eight and ten datasets, respectively.

The performance of MOEA on BM-Set2 (see Table 7) was compared with Gibbs sampler [32], TEPITOPE [35], SVRMHC [8], ARB [10] and NetMHCII [11]. Each allele dataset was subjected to five-fold cross-validation and the results are given in Table 8. The present method shows comparable or superior performance on majority of allele datasets compared to Gibbs sampler, SVRMHC, TEPITOPE, and NetMHCII. A fair comparison of ARB method cannot be drawn because the method has been trained on quantitative data obtained from IEDB [10].

Discussion

We proposed two approaches using MOEA for deriving motifs (1) when the information of only the binders and non-binders are known (i.e., self-discovery) and (2) when, in addition, the information of experimentally (wet-lab) determined motifs are available (i.e., guided-discovery).

Since I-A^{g7} molecule is known to bind to a large number of peptides of low affinity and appears to be a promiscuous binder, the prediction of peptides binding to I-A^{g7} molecule has been nontrivial. This has lead to the definition of a number of suboptimal consensus motifs specific to the datasets. MOEA derived motifs had superior generalization capabilities to those derived with MEME and RANKPEP techniques as well as to the experimentally determined motifs on other datasets. The performances evaluated on two benchmark datasets indicate that the

Table 3: Validation of I-A^{g7} experimental motifs

Experimental Motif	AUC value						
	Datasets						
	Reizis	Harrison	Gregori	Latek	Corper	MHCPEP	Yu
m(Reizis)	0.95	0.68	0.74	0.95	0.50	0.59	0.48
m(Harrison)	0.75	0.88	0.69	0.64	0.53	0.72	0.33
m(Gregori)	0.64	0.68	0.71	0.73	0.40	0.64	0.61
m(Latek)	0.66	0.72	0.80	0.95	0.64	0.52	0.75
m(Rammensee	0.49	0.64	0.76	0.82	0.60	0.48	0.43
m(Reich)	0.55	0.64	0.69	0.58	0.56	0.47	0.50
m(Amor)	0.69	0.54	0.66	0.70	0.56	0.66	0.40

Performance measured by AUC of experimentally determined I-A^{g7} motifs on their own datasets and other experimental datasets.

Table 4: Performance of I-A^{g7} MOEA derived motifs

MOEA-derived Motifs	AUC value						
	Datasets						
	Reizis	Harrison	Gregori	Latek	Corper	MHCPEP	Yu
self-discovery	0.75	0.75	0.77	0.93	0.70	0.75	0.75
guided-discovery	0.77	0.74	0.81	0.83	0.72	0.77	0.71

Seven-fold cross-validation accuracies of MOEA derived motifs on training dataset.

Table 5: Description of peptides in BM-Set1

BM-Set1	Original		NR	
DRB1*0401	Binders	Non-binders	Binders	Non-binders
Set1	694	323	248	283
Set2	381	292	161	255
Set3a	373	217	151	204
Set3b	279	216	128	197
Set4a	323	323	120	283
Set4b	292	292	120	255
Set5a	70	47	65	45
Set5b	48	37	47	37
Southwood	16	6	15	6
Geluk	22	83	19	80

The number of binders and non-binders in the original and non-redundant (NR) datasets in BM-Set1.

Table 6: Comparison of performance on BM-Set1

Dataset		AUC				
		†SVRMHC	Gibbs	ARB	TEPITOPE	MOEA
Original	set1	0.711	0.799	0.666	0.760	0.760
	set2	0.652	0.766	0.653	0.736	0.765
	set3a	0.626	0.740	0.652	0.730	0.733
	set3b	0.618	0.751	0.666	0.750	0.752
	set4a	0.706	0.788	0.668	0.748	0.748
	set4b	0.664	0.770	0.661	0.748	0.770
	set5a	0.553	0.604	0.539	0.653	0.777
	set5b	0.606	0.621	0.579	0.679	0.748
	Southwood	0.912	0.862	0.514	0.490	0.784
	Geluk	0.697	0.723	0.682	0.710	0.786
NR	set1	0.619	0.673	0.572	0.594	0.587
	set2	0.581	0.665	0.640	0.653	0.685
	set3a	0.578	0.598	0.600	0.598	0.660
	set3b	0.577	0.692	0.669	0.699	0.713
	set4a	0.597	0.671	0.575	0.573	0.599
	set4b	0.577	0.669	0.651	0.655	0.690
	set5a	0.544	0.601	0.536	0.646	0.790
	set5b	0.593	0.610	0.572	0.671	0.743
	Southwood	0.917	0.850	0.671	0.505	0.770
	Geluk	0.655	0.697	0.510	0.670	0.768

Comparison of AUC values of the BM-Set1 (DRB1*0401). †These values are based on smaller dataset sizes as SVRMHC didn't predict values for some of the peptides. The values from the Gibbs sampler were estimated from the matrix provided by the authors in [32].

present MOEA based algorithm is applicable in deriving motifs on other class II MHC alleles as well.

The likelihood of finding an optimal motif by MOEA is higher than by a local or greedy search because of the stochastic nature of EA. The proposed approach learns from the characteristics of both binders and non-binders in the training set whereas other methods use information only from binders to determine motifs [27,32]. Moreover, ranges of the parameters involved in MOEA are known, so the parameters of the fitness functions are quickly estimated in a few cross-validation runs. Furthermore, unlike the earlier methods, the present method does not rely on any prior information such as anchor positions to obtain an alignment, prior distributions, etc., [8,9]. Given sufficient data samples representing both binders and non-binders, the method could be applicable to find motifs in other types of molecules. A future direction of this research would be to integrate additional information such as peptide length [69] and PFR [70] as such information has been shown to have the potential to enhance motif detection [11,69]. This would lead to further improvement of the performance of the present algorithm.

Table 7: Description of peptides in BM-Set2

Type	Allele	Binders	Non-binders
Mouse	I-Ab	43	33
	I-Ad	56	286
	I-As	35	91
HLA	DRB1-0101	920	283
	DRB1-0301	65	409
	DRB1-0401	209	248
	DRB1-0404	74	94
	DRB1-0405	88	83
	DRB1-0701	125	185
	DRB1-0802	58	116
	DRB1-0901	47	70
	DRB1-1101	95	264
	DRB1-1302	101	78
	DRB1-1501	188	177
	DRB4-0101	74	107
	DRB5-0101	112	231

The number of binders and non-binders in each of the dataset in BM-Set2. The datasets in BM-Set2 were obtained from [77]. The DRB3-0101 allele dataset was excluded from the performance comparison due to significant imbalance in the dataset (3 binders and 99 non-binders).

Even though EAs are generally known to be computationally intensive, training for derivation of scoring matrices can be performed off-line and the prediction engines can be provided through web services. As seen in Tables 6 and 8, a single method does not always perform well on all types of allele datasets. Nevertheless, the present method showed higher accuracy in detecting motifs on majority of MHC alleles in the benchmark datasets. Therefore, we

Table 8: Comparison of Performance on BM-Set2

Type	Allele	AUC					
		SVRMHC	Gibbs	ARB	TEPITOPE	NetMHCII	MOEA
Mouse	I-A^b	-	-	0.662	-	0.908	0.919
	I-A^d	-	-	0.819	-	0.818	0.855
	I-A^s	-	-	-	-	0.898	0.889
HLA	DRB1-0101	0.623	0.676	0.666	0.647	0.716	0.651
	DRB1-0301	-	0.722	0.799	0.734	0.765	0.778
	DRB1-0401	0.739	0.759	0.737	0.754	0.758	0.725
	DRB1-0404	-	0.743	0.788	0.829	0.785	0.786
	DRB1-0405	0.701	0.724	0.724	0.790	0.735	0.756
	DRB1-0701	-	0.695	0.749	0.768	0.787	0.735
	DRB1-0802	-	0.721	0.803	0.769	0.756	0.773
	DRB1-0901	-	0.734	0.711	-	0.775	0.712
	DRB1-1101	-	0.715	0.727	0.710	0.734	0.759
	DRB1-1302	-	0.716	0.917	0.720	0.818	0.820
	DRB1-1501	0.730	0.672	0.792	0.726	0.736	0.743
	DRB4-0101	-	0.742	0.800	-	0.736	0.759
	DRB5-0101	0.649	0.618	0.677	0.653	0.664	0.660

Comparison of AUC values from five-fold cross-validation of allele datasets given in BM-Set2. "-" indicates that the allele is unavailable for testing with the respective prediction method.

believe that MOEA-based methods could provide a general framework for efficiently determining motifs in a wide range of MHC molecules.

In immunology, accuracy and speed in predicting binding peptides is of paramount importance. Computationally predicted binders do subsequently need to be validated with wet-lab experiments. By using computational predictions as an initial step, high cost involved in initial screening and time-consuming clinical testing can be significantly reduced. Towards this end, the proposed MOEA methods present a promising way to predict peptides that bind to MHC class II alleles including promiscuous and low affinity peptide binders.

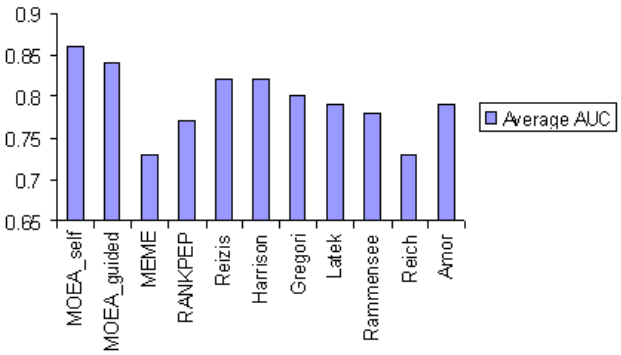

Figure 1
Comparison of Performances. Comparison of performance of MOEA based algorithms – self-discovery and guided-discovery – against MEME, RANKPEP, and experimental motifs on the balanced I-A^{g7} test datasets (the performance was averaged over 25 test datasets)

Conclusion

We present two MOEA-based algorithms for finding motifs, one for self-discovery and the other for guided-discovery by experimentally determined motifs, and thereby predicting binding peptides to I-A^{g7} molecule. Our experiments show that the proposed MOEA-based algorithms are better than earlier methods in predicting binding sites not only on I-A^{g7} but also on most alleles of class II MHC benchmark datasets. This demonstrates the applicability of our methods to find binding motifs in a wide range of MHC alleles.

Methods

Datasets

Several I-A^{g7} datasets were extracted from literature [40-43,62-64] and from Brusic, V.(unpublished data). The numbers of binders and non-binders in each dataset are given in Table 1. The datasets consist of short peptides ranging from 9–30aa in length. Their binding affinities had been experimentally determined by independent studies and classified as binders or non-binders based on IC_{50} values according to the following scheme [41]: good binder (IC_{50} = 100 nM); weak binder (IC_{50} = 2000 nM); non-binder (IC_{50} = 50000 nM). The datasets in [40-43,62-64] were combined into a single training dataset and curated by removing duplicates and redundancy as follows: if a binder is a subsequence of another binder sequence, the longer binder sequence is discarded; if a

non-binder is a subsequence of another non-binder, the shorter subsequence is discarded. Let the curated whole dataset be referred to as *training* dataset here onwards and it be denoted by $D = \{(x_i, v_i): i = 1, 2, ... N\}$ where N is the number of total peptide sequences and x_i is the i-th peptide sequence with the label $v_i \in \{b, nb\}$ indicating whether the sequence x_i is a binder (b) or a non-binder (nb). The number of peptides in the training set $N = 438$ in which the number of binders $N_b = 304$ and the number of non-binders $N_{nb} = 134$.

The set of experimentally validated I-A^{g7}motifs [38-44] derived largely from uncorrelated datasets [40-43] was extracted and is illustrated in Table 1 with the distribution of binders and non-binders in each dataset. Table 2 illustrates an experimentally validated motif of I-A^{g7} reported by Reizis *et al* [40]. Experimental motifs are described by the anchor positions and binding affinities of amino acids of the motif. The residues which contribute significantly to the peptide binding are called primary anchor residues and positions they reside are called anchor positions. An amino acid occupying a specific position within a motif is characterized as well tolerated, weakly tolerated, or non-tolerated based on its involvement in the binding process.

An independent dataset was generated from binders of Stratmann dataset [46], consisting of a diverse set of I-A^{g7} binding peptides with their binding affinities, to find the test accuracies in predicting binders and non-binders. The Stratmann dataset was balanced with randomly generated 9-mer non-binders so that for testing dataset, $N_b = N_{nb} = 112$.

Binding Score Matrix

A k-mer motif of amino acids is characterized by a PSSM $Q = \{q_{ia}\}_{k \times 20}$ where q_{ia} denotes the binding strength of the site i when it is occupied by amino acid a. The binding score of a putative motif is computed by adding the binding scores assigned to each amino acid at the respective positions. The binding score indicates the likelihood of the motif binding to the molecule. The binding score s_i of sequence $x_i = (x_{i,1}, x_{i,2}, ...x_{i,n})$ of length n is determined by the maximum value of binding scores computed for all k-mer subsequences in x_i:

$$s_i = \max_j \{s_{ij} : j = 1, 2, \cdots n - k + 1\} \quad (1)$$

where s_{ij} denotes the binding score of the subsequence beginning at location j of the sequence i, which is given by

$$s_{ij} = \sum_{l=1,2,\cdots k} q_{(j+l),} x_{i(j+l)} \quad (2)$$

and assuming that only one motif instance exists in every sequence, the location j^* of the motif is given by

$$j^* = \arg\max_j \{s_{ij} : j = 1, 2, \cdots n - k + 1\} \quad (3)$$

That is, the most likely motif instance of sequence x_i, say m_i, is given by the sequence $m_i = (x_{ij^*} \cdot x_{ij^*+1}, ... x_{ij^*+k-1})$.

Self-discovery of Motif

We derive a consensus motif from the training dataset which consists of peptides from several experiments and of varying lengths. The positions of binding cores within the peptides are unknown. The elements of the PSSM are represented as $20k$-tuples $(q_{ia}, : i = 1, ... k; a \in \Omega)$ where Ω represents the amino acid alphabet. Each element in the k-tuple is converted to a real number representation using a binary word of size θ so that $q_{ia} \in [0, 2^\theta - 1]$. The k-mer motif is therefore represented by an individual of $20k\theta$ long string in the EA. Let the population at t-th iteration of the evolution is denoted by $q(t) = \{q_1(t), q_2(t), ... q_M(t)\}$ where $q_j(t)$ represents an individual in a population of size M.

The fitness function is designed to arrive at an optimal consensus of the motif, by using the training dataset. A solution is evaluated based on its ability to maximize the accuracies in identifying true binders (TP) and true non-binders (TN) as well as to widen the gap between the total score for binders and non-binders. This is achieved by two fitness functions: f_1 to minimize the sum of false positives (FP) and false negatives (FN), and f_2 to minimize the ratio between the average cumulative scores of non-binders and binders:

$$f_1 = FN + \kappa_1 FP \quad (4)$$

$$f_2 = \frac{N_b}{N_{nb}} \frac{\sum_{i=1}^{N} s(m_i)\delta(v_i = nb)}{\sum_{i=1}^{N} s(m_i)\delta(v_i = b)} \quad (5)$$

Eqs. (4) and (5) are minimized and subjected to following two constraints:

$$\frac{FP}{N_{nb}} \leq \frac{1}{\alpha_1} \quad (6)$$

$$\frac{FN}{N_b} \leq \frac{1}{\alpha_2} \quad (7)$$

where $s(m_i)$ denotes the score computed for the most likely motif instance m_i of sequence x_i of the training dataset, and Kronecker δ is one when the argument is satisfied and otherwise is zero. N_b and N_{nb} are the total counts of

binders and non-binders in the dataset. The constant κ_1 ($>N_b/N_{nb}$ for $N_b > N_{nb}$, or vice versa) was empirically determined to minimize the number of false positives. The two parameters α_1 ($<<N_{nb}$) and α_2 ($<<N_b$) are set to minimize FP and FN rates, respectively. If none of the individuals satisfies the above constraints, MOEA reports no feasible solution. Given the training set, a few trial runs with different initializations are necessary to determine the best values of α_1 and α_2.

Scoring of Experimental Motifs

The description of an experimental k-mer motif conveys three kinds of information at each site: (1) the amino acid occupied, (2) the tolerance level of the amino acid, and (3) the strength of binding. Let us denote a k-mer motif validated in experiment "e" by $m(e)$ and the tolerance level of the residue at site j by ρ_j where $\rho_j \in$ {well, weak, unknown, non – tolerated}. The binding strength of site j is expressed by $\sigma_j \in$ {primary – anchor, secondary – anchor, other}. Then, the binding score for a k-mer experimental motif is given by

$$s(m(e)) = \sum_{j=1}^{k} \rho_j \cdot \sigma_j \quad (8)$$

Guided-discovery of Motif

In this algorithm, we assume that experimentally determined motifs are available along with the experimental datasets. An MOEA is proposed to determine a motif closer to experimental motifs. An objective function f_3 is proposed to best represent the characteristics of the motif that is close to the knowledge embedded in the experimental motifs:

$$f_3 = \sum_{e} \left| \hat{Q} - Q(m(e)) \right| \quad (9)$$

where $\hat{Q}$ denotes the estimated PSSM of the motif. We use the same objective function in Eq. (4) to accurately predict binders of the training dataset. The MOEA minimizes the objective functions given in Eqs. (4) and (9), subjected to the two constraints given in Eqs. (6) and (7). The summation in Eq. (9) is taken over all the experimental motifs and $|\hat{Q} - Q(m(e))|$ is the sum of squares of differences between individual elements of weight matrices $\hat{Q}$ and $Q(m(e))$. The knowledge of the experimental motif is incorporated to the consensus motif adaptively with the distance function used in f_3. Further, the fitness f_1 optimizes the specificity and sensitivity of the prediction of binders.

The elements in the PSSM of experimental motifs are set to values within the same range [0, 2^{θ}-1] as before. The following procedure is adopted to determine the elements of $Q(m(e))$: a well tolerated amino acid at an anchor position of the motif receives the highest possible score of 2^{θ}-1; the lowest score of zero is assigned to a non-tolerated residue; weakly tolerated residues and residues at secondary anchor positions receive of $(2^{\theta}-1)/2$; and all the other unknown positions receives a score of $(2^{\theta}-1)/3$.

Performance Comparison

The binding scores of I-A^{g7} experimental motifs were computed using Eq. (8) by assigning the following values for binding strengths: primary = 4, secondary = 2, and others = 1, and for anchor positions: well = 4, weak = 2, non-tolerated = -4, and unknown = 0. The experimentally determined motifs were used with peptide data in the guided-discovery of motifs.

We used AUC to compare performance of the proposed methods with earlier approaches [28,34] and experimental motifs [38-44]. Whether a peptide is a binder or a non-binder is determined by a threshold of the binding score. By varying this threshold, the ROC curve was plotted, from which AUC value was obtained. A comparison of performances of the methods is given in Figure 1.

In order to compare to the MEME method, only binders in the I-A^{g7} training set were submitted to MEME motif discovery tool at the prediction server [71]. The motif of 9-mer length was obtained with the following options: zero or one motif per sequence, minimum and maximum width = 9. The performance accuracy of RANKPEP approach on the testing dataset was carried out by uploading the dataset to the online prediction server at [72] with a 4% binding threshold [34].

Benchmark Datasets

The proposed self-discovery approach was tested on BM-Set1, i.e., HLA-DRB1*0401, which consists of one training set and 10 testing datasets and had been earlier used to benchmark a number of motif finding algorithms [25,26,32,73]. The performance of MOEA was compared with earlier methods [8,10,32,35].

The training set consisting of binders and non-binders was assembled as follows: an ensemble of 532 unique binding peptides were extracted from SYFPEITHI [44] and MHCPEP [63] databases and a set of 177 unique non-binders were extracted from the MHCBN database [20]. The datasets were pre-processed by removing peptides that did not allow a hydrophobic residue at P1 position of all putative 9-mer binding cores and unnatural peptides containing more than 75% alanine [32]. The preprocessed

binder set has 456 unique peptides with a length distribution ranging from 9 to 30 amino acid residues.

Of the 10 testing datasets, 8 datasets were taken from the MHC-bench as described in [74]. The other 2 datasets were extracted from experiments described by Southwood [75] and Geluk [76]. An affinity of (IC_{50} = 1000 nM) was taken as the threshold for peptide binding as described in [75]. Homology reduction had been carried out on all datasets in order to reduce the chances of over-fitting due to the redundancy of datasets. The peptides in the non-redundant (NR) datasets had sequence similarities less than 90%. The number of binders and non-binders in the original and NR datasets are given in Table 5.

We tested our method on BM-Set2 comprising of 3 mouse alleles and 13 HLA alleles made available at [77]. These quantitative peptide datasets had been extracted from the IEDB at [78]. The number of binders and non-binders in each dataset is given in Table 7. The DRB3-0101 allele dataset was excluded from the benchmark dataset because of the significant imbalance between binders and non-binders (3 binders and 99 non-binders). With this dataset, we compared our method with [8,10,11,32,35].

Parameters of MOEA

The range of positional scores was set with $\theta = 7$. For each run of MOEA, the population size M = 500, crossover probability p_c = 0.9, and mutation probability p_m = 0.005 were used. The process was terminated after 300 generations as no significant improvement in the convergence was observed during the experimental trial sessions. The parameters of the fitness functions were empirically determined for optimum performance within the following ranges: κ_1 = 1~2.5, α_1 = 5.0–6.0, and α_2 = 1.0–2.0. The parameters κ_1 = 2.5, α_1 = 6.0, and α_2 = 2.0 were found to work well empirically for both datasets.

Authors' contributions

MR and VB conceived the study; MR designed experiments and performed computational analysis; MR, BS, VB and LF wrote the manuscript. All authors read and corrected the manuscript.

Additional material

Additional file 1

MOEA derived matrices on I-A^{g7} dataset. The two PSSM derived by using MOEA self-discovery and guided-discovery approaches are given in the Additional file 1.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-459-S1.pdf]

Additional file 2

Motif logos obtained for I-A^{g7} from MOEA derived matrices. Figure 1 and Figure 2 illustrate motif logos derived from the alignments obtained from the MOEA guided-discovery and self-discovery approaches. The web server [79] was used to generate the motif logos as described in [68].
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-459-S2.pdf]

Acknowledgements

The authors would like to thank Dr. Tim Oliver for proof reading the manuscript. We are also grateful to the anonymous reviewers whose comments significantly improved the paper.

References

1. Stern LJ, Wiley DC: **Antigenic peptide binding by class I and class II histocompatibility proteins.** *Behring Inst Mitt* 1994:1-10.
2. Hammer J, Bono E, Gallazzi F, Belunis C, Nagy Z, Sinigaglia F: **Precise prediction of major histocompatibility complex class II-peptide interaction based on peptide side chain scanning.** *J Exp Med* 1994, **180(6):**2353-2358.
3. Rammensee HG, Friede T, Stevanoviic S: **MHC ligands and peptide motifs: first listing.** *Immunogenetics* 1995, **41(4):**178-228.
4. Mamitsuka H: **Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models.** *Proteins* 1998, **33(4):**460-474.
5. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG: **Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules.** *Nature* 1991, **351(6324):**290-296.
6. Ruppert J, Sidney J, Celis E, Kubo RT, Grey HM, Sette A: **Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules.** *Cell* 1993, **74(5):**929-937.
7. Bouvier M, Wiley DC: **Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules.** *Science* 1994, **265(5170):**398-402.
8. Wan J, Liu W, Xu Q, Y R, Flower DR, Li T: **SVRMHC prediction server for MHC-binding peptides.** *BMC Bioinformatics* 2006, **7:**463.
9. Doytchinova IA, Flower DR: **Towards the insilico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction.** *Bioinformatics* 2003, **19(17):**2263-2270.
10. Bui H, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton K, Mothé BR, Chisari FV, Watkins DI, Sette A: **Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications.** *Immunogenetics* 2005, **57(5):**304-314.
11. Nielsen M, Lundegaard C, Lund O: **Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method.** *BMC Bioinformatics* 2007, **8(238):**.
12. Bisset L, Fierz W: **Using a neural network to identify potential HLA-DR1 binding sites within proteins.** *J Mol Recognition* 1994, **6:**41-48.
13. Brusic V, Rudy G, Harrison LC: **Prediction of MHC binding peptides using artificial neural networks.** In *Complex Systems: Mechanism of Adaptation* Edited by: Stonier R, Yu XS. Amsterdam: IOS Press; 1994:253-260.
14. Adams HP, Koziol JA: **Prediction of binding to MHC class I molecules.** *J Immunol Methods* 1995, **185(2):**181-190.
15. Gulukota K, Sidney J, Sette A, DeLisi C: **Two complementary methods for predicting peptide binding major histocompatibility complex molecules.** *J Mol Biol* 1997, **267:**1258-1267.
16. Burden FR, Winkler DA: **Predictive Bayesian neural network models of MHC class II peptide binding.** *J Mol Graph Model* 2005, **23(6):**481-489.
17. Noguchi H, Kato R, Hanai T, Matsubara Y, Honda H, Brusic V, Kobayashi T: **Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules.** *J Biosci Bioeng* 2002, **94(3):**264-270.

18. Donnes P, Elofsson A: **Prediction of MHC class I binding peptides, using SVMHC.** *BMC Bioinformatics* 2002, **3:**25.
19. Zhao Y, Pinilla C, Valmori D, Martin R, Simon R: **Application of support vector machines for T-cell epitopes prediction.** *Bioinformatics* 2003, **19(15):**1978-1984.
20. Bhasin M, Singh H, Raghava GPS: **MHCBN: A comprehensive database of MHC binding and non-binding peptides.** *Bioinformatics* 2003, **19:**665-666.
21. Salomon J, Flower DR: **Predicting Class II MHC-Peptide binding: a kernel based approach using similarity scores.** *BMC Bioinformatics* 2006, **7:**551.
22. Takahashi H, Honda H: **Prediction of peptide binding to major histocompatibility complex class II molecules through use of bossted fuzzy classifier with SWEEP operator method.** *Bioscience and Bioengineering* 2006, **101(2):**137-141.
23. Mallios RR: **Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm.** *Bioinformatics* 1999, **15(6):**432-439.
24. Mallios RR: **Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm.** *Bioinformatics* 2001, **17(10):**942-948.
25. Murugan N, Dai Y: **Prediction of MHC class II binding peptides based on an iterative learning model.** *Immunome Res* 2005, **1(6):**10.
26. Karpenko O, Shi J, Dai Y: **Prediction of MHC class II binders using the ant colony search strategy.** *Artif Intell Medicine* 2005, **35(1–2):**47-56.
27. Bailey TL, Elkan C: **Unsupervised learning of multiple motifs in biopolymers using expectation maximization.** *Machine Learning* 1995, **21:**51-80.
28. Bailey TL, Charles E: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** In *Second International Conference on Intelligent Systems for Molecular Biology* AAAI Press, Menlo Park, California; 1994:28-36.
29. Neuwald AF, Liu JS, Lawrence CE: **Gibbs motif sampling: detection of bacterial outer membrane protein repeats.** *Protein Sci* 1995, **4(8):**1618-1632.
30. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y: **A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes.** *J Comput Biol* 2002, **9(2):**447-464.
31. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262(5131):**208-214.
32. Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, Buus S, Brunak S, Lund O: **Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach.** *Bioinformatics* 2004, **20(9):**1388-1397.
33. Reche PA, Glutting JP, Reinherz EL: **Prediction of MHC class I binding peptides using profile motifs.** *Hum Immunol* 2002, **63(9):**701-709.
34. Reche PA, Glutting JP, Zhang H, Reinherz EL: **Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles.** *Immunogenetics* 2004, **56(6):**405-419.
35. Sturniolo T, Bono E, Jiayi D, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti M, Sinigaglia F, Hammer J: **Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices.** *Nature Biotech* 1999, **17(6):**555-561.
36. Brusic V, Schonbach C, Takiguchi M, Ciesielski V, Harrison LC: **Application of genetic search in derivation of matrix models of peptide binding to MHC molecules.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**75-83.
37. Rajapakse M, Schmidt B, Brusic V: **Multi-Objecitve Evolutionary Algorithm for Discovering Peptide Binding Motifs.** In *Applications of Evolutionary Computing Volume 3907*. Lecture Notes in Computer Science, Springer; 2006:149-158.
38. Reich EP, von Grafenstein H, Barlow A, Swenson KE, Williams K, Janeway CA Jr: **Self peptides isolated from MHC glycoproteins of non-obese diabetic mice.** *J Immunol* 1994, **152(5):**2279-2288.
39. Amor S, O'Neill JK, Morris MM, Smith RM, Wraith DC, Groome N, Travers PJ, Baker D: **Encephalitogenic epitopes of myelin basic protein, proteolipid protein, myelin oligodendrocyte glycoprotein for experimental allergic encephalomyelitis induction in Biozzi ABH (H-2Ag7) mice share an amino acid motif.** *J Immunol* 1996, **156(8):**3000-3008.
40. Reizis B, Eisenstein M, Bockova J, Konen-Waisman S, Mor F, Elias D, Cohen IR: **Molecular characterization of the diabetes-associated mouse MHC class II protein, I-Ag7.** *Int Immunol* 1997, **9(1):**43-51.
41. Harrison LC, Honeyman MC, Trembleau S, Gregori S, Gallazzi F, Augstein P, Brusic V, Hammer J, Adorini L: **A peptide-binding motif for I-A(g7), the class II major histocompatibility complex (MHC) molecule of NOD and Biozzi AB/H mice.** *J Exp Med* 1997, **185(6):**1013-1021.
42. Latek RR, Suri A, Petzold SJ, Nelson CA, Kanagawa O, Unanue ER, Fremont DH: **Structural basis of peptide binding and presentation by the type I diabetes-associated MHC class II molecule of NOD mice.** *Immunity* 2000, **12(6):**699-710.
43. Gregori S, Bono E, Gallazzi F, Hammer J, Harrison LC, Adorini L: **The motif for peptide binding to the insulin-dependent diabetes mellitus-associated class II MHC molecule I-Ag7 validated by phage display library.** *Int Immunol* 2000, **12(4):**493-503.
44. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S: **SYFPEITHI: database for MHC ligands and peptide motifs.** *Immunogenetics* 1999, **50(3–4):**213-219.
45. Carrasco-Marin E, Kanagawa O, Unanue ER: **The lack of consensus for I-A(g7)-peptide binding motifs: is there a requirement for anchor amino acid side chains?** *Proc Natl Acad Sci USA* 1999, **96(15):**8621-8626.
46. Stratmann T, Apostolopoulos V, Mallet-Designe V, Corper AL, Scott CA, Wilson IA, Kang AS, Teyton L: **The I-Ag7 MHC class II molecule linked to murine diabetes is a promiscuous peptide binder.** *J Immunology* 2000, **165(6):**3214-3225.
47. Carrasco-Marin E, Shimizu J, Kanagawa O, Unanue ER: **The class II MHC I-Ag7 molecules from non-obese diabetic mice are poor peptide binders.** *J Immunol* 1996, **156(2):**450-458.
48. Suri A, Vidavsky I, van der Drift K, Kanagawa O, Gross ML, Unanue ER: **In APCs, the autologous peptides selected by the diabetogenic I-Ag7 molecule are unique and determined by the amino acid changes in the P9 pocket.** *J Immunol* 2002, **168(3):**1235-1243.
49. Beiko RG, Charlebois RL: **GANN: genetic algorithm neural networks for the detection of conserved combinations of features in DNA.** *BMC Bioinformatics* 2005, **6(1):**36.
50. Fogel GB, Weekes DG, Varga G, Dow ER, Harlow HB, Onyia JE, Su C: **Discovery of sequence motifs related to coexpression of genes using evolutionary computation.** *Nucleic Acids Res* 2004, **32(13):**3826-3835.
51. Liu F, Tsai J, Chen R, Chen S, Shih S: **FMGA: Finding Motifs by Genetic Algorithm.** *IEEE BIBE* 2004.
52. Lo N, Changchien S, Chang Y, Lu T: **Human promoter prediction based on sorted consensus sequence patterns by genetic algorithms.** *Intl congr on Biological and Medical Engineering* 2002:111-112.
53. Corne D, Meade A, Sibly R: **Evolving Core Promoter Signal Motifs.** *IEEE Congress on Evolutionary Computation* 2001:1162-1169.
54. Fogel G, Corne D: **Evolutionary Computation in Bioinformatics.** Morgan Kaufman publishers; 2003.
55. Mitchell M: **An Introduction to Genetic Algorithms.** MIT press; 1999.
56. Deb K: **Multi-Objective Optimization Using Evolutionary Algorithms.** Wiley publishers; 2001.
57. Holland J: **Adaptation in Natural and Artificial Systems.** Ann Arbor, MI: University of Michigan Press; 1975.
58. Deb K, Pratap A, Agrawal S, Meyarivan T: **A Fast and Elitist Multiobjective Genetic Algorithm:NSGA-II.** *IEEE Trans on Evolutionary Computation* 2002, **6(2):**182-197.
59. Zitzler E, Thiele L: **Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength of Pareto Approach.** *IEEE Trans on Evolutionary Computation* 1999, **3:**257-271.
60. Knowles JD: **Approximating the Nondominant front using the Pareto Archived evolution strategy.** In *Evolutionary Computation Volume 8*. Issue Summer MIT Press; 2000:49-172.
61. Fonseca C, Fleming PJ: **Genetic Algorithms for Multiobjective Optimization: Formulation, discussion and generalization.** In *the fifth Intl conference on Genetic Algorithms* San Mateo, CA: Morgan Kauffman; 1993:416-423.
62. Corper AL, Stratmann T, Apostolopoulos V, Scott CA, Garcia KC, Kang AS, Wilson IA, Teyton L: **A Structural Framework for**

Deciphering the Link Between I-Ag7 and Autoimmune Diabetes. *Science* **288(5465):**505-511. 21 April 2000
63. Brusic V, Rudy G, Harrison LC: **MHCPEP, a database of MHC-binding peptides: update 1997.** *Nucleic Acids Res* 1998, **26(1):**368-371.
64. Yu B, Gauthier L, Hausmann DH, Wucherpfennig KW: **Binding of conserved islet peptides by human and murine MHC class II molecules associated with susceptibility to type I diabetes.** *Eur J Immunol* 2000, **30(9):**2497-2506.
65. Webb A: *Statistical Pattern Recognition* 2nd edition. John Wiley & Sons; 2002.
66. Swets JA: **Measuring the accuracy of diagnostic systems.** *Science* 1988, **240(4857):**1285-1293.
67. Schueler-Furman O, Altuvia Y, Sette A, Margalit H: **Structure-based prediction of binding peptides to MHC class I molecules: Application to a broad range of MHC alleles.** *Protein Sci* 2000, **9:**1838-1846.
68. Schneider D, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Research* 1990, **18(20):**6097-6100.
69. Chang ST, Ghosh D, Kirschner DE, Linderman JJ: **Peptide length-based prediction of peptide-MHC class II binding.** *Bioinformatics* 2006, **22(22):**2761-2767.
70. Godkin AJ, Smith KJ, Willis A, Tejada-Simon MV, Zhang J, Elliott T, Hill AVS: **Naturally Processed HLA Class II Peptides Reveal Highly Conserved Immunogenic Flanking Region Sequence Preferences That Reflect Antigen Processing Rather Than Peptide-MHC Interactions.** *Immunology* 2001, **166(11):**6720-6727.
71. **MEME** [http://meme.sdsc.edu/meme/]
72. **RANKPEP** [http://bio.dfci.harvard.edu/Tools/rankpep.html]
73. Bhasin M, Raghava GP: **SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence.** *Bioinformatics* 2004, **20(3):**421-423.
74. **MHCBench** [http://www.imtech.res.in/raghava/mhcbench]
75. Southwood S, Sidney J, Kondo A, del Guercio M, Appella E, Hoffman S, Kubo RT, Chestnut R, Grey HM, Sette A: **Several common HLA-DR types share largely overlapping peptide binding repertoires.** *Immunology* 1998, **160:**3363-3373.
76. Geluk A, van Meijgaarden K, Schloot N, Drijfhout J, Ottenhoff T, Roep B: **HLA-DR binding analysis of peptides from islet antigens in IDDM.** *Diabetes* 1998, **47(1584–1600):**.
77. **NetMHCII** [http://www.cbs.dtu.dk/services/NetMHCII]
78. **IEDB** [http://www.immuneepitope.org]
79. **Weblogo** [http://weblogo.berkeley.edu/]